Saisir le sens dans les deux sens

Mahé Ben Hamed, Damon Mayaffre

Université de Liège / Facultés Universitaires Saint-Louis - Bruxelles

JADT 2012

11es Journées internationales d'analyse statistique des données textuelles

Actes - Proceedings

Anne Dister
Dominique Longrée
Gérald Purnelle (éds)

Saisir le sens dans les deux sens: Exploration de la portée interprétative de l'énergie et de la disponibilité

Mahé Ben Hamed[1], Damon Mayaffre[2]

[1] BCL – Université Nice-Sophia Antipolis, CNRS – mbenhamed@unice.fr
[2] BCL – Université Nice-Sophia Antipolis, CNRS – damon.mayaffre@unice.fr

Abstract

Co-occurrence is an asymmetric relationship: each word forming the pair brings with it its own semantic load that comes from its own network of semantic associations. Energy and availability (Luong *et al.*, 2010) measure this asymmetry and we suggest here that they grasp quantitatively the notion of semantic load. Re-drafting them in a probabilistic framework to make them comparable between themselves and across words, we explore their interpretative value in discourse analysis on a corpus of French political speeches. We also introduce the notion of discursive performance that articulates both notions into one, and we explore the added value it brings to availability diagnostics for the purpose of identifying the modalities of semantic contextualization of a word by a given speaker.

Résumé

Les relations co-occurrentielles sont des relations asymétriques : chaque mot apporte à la paire une charge sémantique qui lui vient de son propre réseau d'associations sémantiques. Les mesures d'énergie et de disponibilité (Luong *et al.*, 2010) quantifient cette asymétrie, et nous proposons ici qu'elles saisissent quantitativement la notion de charge sémantique. En les reformulant dans un cadre probabiliste afin de les rendre comparables entre elles et d'un mot à l'autre, nous explorons leur portée interprétative en analyse du discours sur un corpus politique français. Nous introduisons également la notion de rendement discursif pour articuler entre elles les deux notions et explorons son utilité pour nuancer le diagnostic de la disponibilité sur la modulation sémantique d'un mot par un locuteur.

Mots-clés : co-occurrence, association asymétrique, charge sémantique, corpus politique, thématisation

1. Introduction

De Firth (1957) hier à Rastier aujourd'hui (2011), la linguistique des corpus textuels reconnait la co-occurrence comme centrale à sa pratique. La co-occurrence constitue la maille textuelle minimale de contextualisation sémantique d'une lexie, et traverse dès lors des problématiques allant de la cohésion textuelle ou de la textualité à la thématisation en discours. La littérature TAL et ADT propose d'ailleurs une multitude de mesures co-occurrrentielles supposées saisir, au sein d'un texte ou d'un corpus, l'attraction existant entre deux mots. La plus connue est

sans doute l'indice *d'information mutuelle* (Church et Hanks, 1991) qui mesure le rapport de vraisemblance entre une attraction préférentielle de A et B et une association qui serait purement aléatoire. Des auteurs comme Viprey (1997), Martinez (2003) ou Heiden (2004) abordent plus finement la notion de co-occurrence au travers d'analyses multidimensionnelles, ou d'une extension de la notion de co-occurrence binaire à celles de co-occurrences multiples, de poly-occurrences ou de profil co-occurrentiel pour rendre compte de la structuration de l'espace sémantique du texte ou du corpus.

Cette littérature, comme le note Luong *et al.* (2010), aborde cependant l'attraction entre deux mots comme une relation symétrique à laquelle les deux termes en présence contribuent de façon identique en force et en nature. Or chaque pôle de la co-occurrence possède une fréquence relative et une distribution propre dans le texte ainsi qu'un réseau spécifique d'associations qu'il entretient avec les autres lexies du texte, et on s'attend dès lors à ce que les deux pôles n'interviennent pas à mesure égale dans la constitution de leur association. Partant de ce constat, Luong *et al.* (2010) proposent un couple notionnel dit d'énergie et de disponibilité, qui caractérise le mot, et qui serait en mesure de décrire de façon polarisée son comportement associatif vis-à-vis des autres mots.

Dans cet article, nous poursuivons l'exploration du potentiel heuristique de ces mesures d'énergie et de disponibilité : **(i)** en proposant une formulation analytique normalisée par la réinscription des notions d'énergie/disponibilité dans un cadre probabiliste (section 2); **(ii)** en explorant la portée interprétative des profils énergie/disponibilité *versus* une représentation synthétique d'un comportement global du mot pour l'ensemble des associations sémantiques considérées. Nous introduisons également la notion de *rendement discursif* qui préserve l'information spécifique à chaque association tout en articulant les deux mesures d'énergie et de disponibilité (section 3) et enfin, **(iii)** en explorant la portée de la disponibilité et de la notion de rendement discursif introduite en section 3 pour la détection de modalités de thématisation d'un mot par un locuteur (section 4).

Dans la perspective discursive qui est la nôtre, l'enjeu est important. On soupçonne en effet des profils lexicaux-sémantiques différents selon les mots, et l'approche polarisée que nous développons ici permet de les mettre en évidence. En travaillant uniquement sur les substantifs afin d'évacuer les questions de dépendance grammaticale, nous montrons la plus-value interprétative qu'il y a à décrire le profil co-occurrentiel d'un mot de façon asymétrique pour distinguer des différences d'usage en discours, et identifier les co-occurrences essentielles à sa thématisation.

2. Reformulation probabiliste de l'énergie et de la disponibilité

2.1. Formulation originale

Luong *et al.* (2010) définissent, pour un mot A et dans sa relation co-occurrentielle avec tout B, son énergie ε et sa disponibilité δ comme : $\varepsilon(A) = \frac{n(AB)}{n(A)}; \delta(A) = \frac{n(AB)}{n(B)}$ (1), ce qui implique nécessairement que: $\varepsilon(A) = \delta(B); \delta(A) = \varepsilon(B)$ (2).En rapportant le nombre de co-occurrences AB au nombre d'occurrences de l'un ou l'autre pôle A ou B, Luong *et al.* (2010) estiment mesurer la part structurante de chacun dans leur attraction relative. Ces mesures sont calculées

pour tout B, et pour tout A, déclinant ainsi,pour chaque mot, des profils en énergie ou en disponibilité qui correspondent à une description polarisée du comportement associatif de tout mot avec tout autre.

Ces mesures d'énergie et de disponibilité sont définies à partir du *nombre d'occurrences* de AB et de A, et non de leur fréquence relative. Or l'espace de co-occurrence de AB n'est pas nécessairement le même que celui d'occurrence de A, et n'est pas non plus nécessairement identique à celui d'une autre co-occurrence AC de A ou d'une toute autre paire DE. Ceci pose un problème de normalisation, et donc de comparabilité : d'abord entre énergie et disponibilité pour un même mot, ensuite, et pour chacune de ces mesures, entre des mots différents. Ajoutons à cela qu'il existe dans la littérature, et implémentés dans les logiciels d'analyse textuelle, différents modes de décompte des co-occurrences (présence/absence de la co-occurrence dans une fenêtre de décompte fixe, ou n(AB) = n(A) x n(B) dans cette même fenêtre, ou encore fenêtre glissante…) qui peuvent amener à comparer, d'un corpus à l'autre et d'un utilisateur à l'autre, des quantités qui ne sont pas comparables.

2.2. Reformulation dans le cadre des probabilités conditionnelles

Afin de normaliser les mesures d'énergie et de disponibilité, il convient tout d'abord de passer d'une formulation en termes de fréquence absolue (nombre d'occurrences) à une formulation en termes de fréquence relative : $\varepsilon(A)=\frac{\upsilon(AB)}{\upsilon(A)};\delta(A)=\frac{\upsilon(AB)}{\upsilon(B)}$ (3), une reformulation qui n'est pas sans rappeler la définition suivante de la probabilité conditionnelle : $p(a|b)=\frac{p(ab)}{p(b)}$; $p(a|b)$ étant la probabilité de *a* sachant *b* (4). Selon l'usage de considérer la fréquence d'occurrence observée comme la probabilité de cette occurrence, nous reformulons les notions d'énergie ε et de disponibilité δ de A dans sa relation co-occurrentielle avec B comme :

$$\begin{cases}\varepsilon(A) = p(A \text{ dans paragraphe}) \times \text{p(B}|\text{A)} \\ \delta(A) = p(B \text{ dans paragraphe}) \times \text{p(A}|\text{B)}\end{cases} \quad (5)$$

Autrement dit, l'asymétrie associative du couple AB est reformulée en terme de *prédictibilité de l'un sachant l'autre* : si A est présent, quelle est la probabilité d'observer également B ? En calculant ceci pour tout B, nous obtenons, comme Luong *et al.* (2010) des *profils* d'énergie et de disponibilité de A. Cette reformulation préserve l'équivalence originale attendue par Luong *et al.* (2010) en (2) et évacue la question de la normalisation et du mode de décompte de la co-occurrence. Elle n'évacue cependant pas la question de la fenêtre de décompte. En effet, l'implémentation pratique que nous proposons pour cette reformulation reste conditionnelle de ce qu'on estime être la fenêtre pertinente pour le décompte des associations. Nous intéressant particulièrement aux associations libres, c'est-à-dire non conditionnées grammaticalement, nous déterminons cette fenêtre comme étant fixe, et correspondant au paragraphe.

2.3. Energie, disponibilité et charge sémantique

Luong *et al.* (2010) suggèrent que l'énergie pourrait mesurer un fonctionnement en langue, là où la disponibilité mesure un fonctionnement en discours. Notre reformulation analytique ne peut à elle seule confirmer cette conclusion, mais détermine elle aussi des prédictions différentes

quant au comportement d'un mot A selon le rôle qui lui est attribué dans la co-occurrence. Le profil énergie d'un mot A prédit la préférence associative relative de A pour chacun des mots considérés dans le profil, en attribuant à A un rôle d'attracteur. Il s'agit, en quelque sorte, de son *potentiel d'amorçage sémantique* dans le champ/espace sémantique couvert par le profil étudié. Le profil disponibilité de A, pour sa part, définit l'espace sémantique probabilisé des attracteurs de A, autrement dit, de son potentiel *à être amorcé* dans le champ sémantique couvert par le profil étudié, que l'on pourrait également qualifier comme étant la *charge sémantique de A*. Du fait de l'équivalence (2), il serait donc possible de lire les profils simplement en termes de charge sémantique : *l'énergie montre comment A confère une charge sémantique à ses associés, et la disponibilité montre comment A obtient la sienne à partir de ses associations sémantiques.*

3. Exploration heuristique de la portée interprétative du couple ε / δ

3.1. Corpus et méthode

Le corpus sur lequel porte notre expérimentation est un corpus contrastif de discours politiques français contemporains prononcés dans le cadre de la campagne électorale 2007. Il comprend 148 discours et 750.000 mots, répartis entre 6 locuteurs (Laguiller, Buffet, Royal, Bayrou, Sarkozy, Le Pen). Seuls les principaux substantifs seront étudiés afin de ne pas introduire dans notre analyse des considérations de dépendance grammaticale, comme celles pouvant exister entre un déterminant et un nom, un nom et un adjectif, etc. Dans ce premier temps de l'analyse, nous nous intéressons à une liste de 75 substantifs sélectionnés sous la double contrainte d'être dans le groupe des 300 à 400 mots les plus fréquents du corpus et communs aux 6 locuteurs.

3.2. Des comportements ε / δ différenciés

Comme Luong *et al.* (2010), les profils constitués par notre reformulation analytique du couple ε / δ diffèrent entre ε et δ pour un même mot A (Fig.1), et de façon peut être plus intéressante d'un point de vue discursif, pour un même mot selon différents locuteurs (Fig. 2).

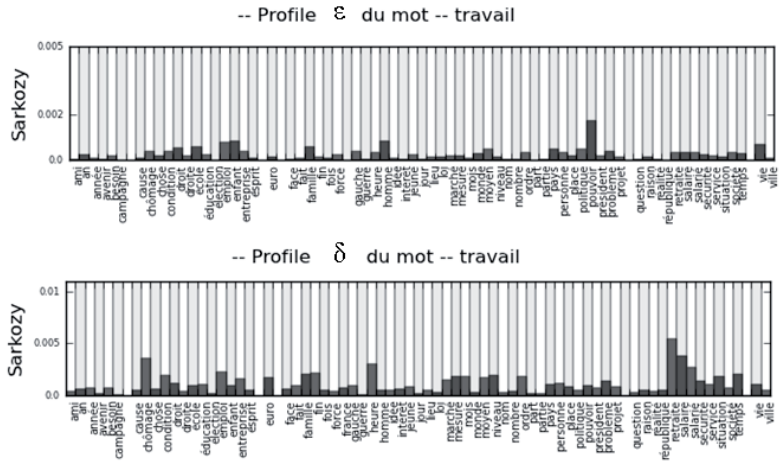

Figure 1 - Energie ε et disponibilité δ du mot 'travail' chez le candidat Sarkozy pour les 75 mots du vocabulaire commun aux 6 candidats. L'ordre des mots est le même (alphabétique) pour les deux profils.

Du fait de l'équivalence notée en (2) et de la relation constatée entre disponibilité, et de ce qu'on qualifie communément (en analyse des données textuelles ou en analyse du discours) de charge sémantique, nous nous intéresserons particulièrement à la notion de disponibilité, puisque toute conclusion sur sa portée interprétative peut être projetée directement en termes (inverses) d'énergie.

Le potentiel heuristique de ces graphiques apparaît évident pour l'analyste du discours. Particulièrement, la distribution des profils de disponibilité de 'travail' chez les différents candidats (Fig.2) montre des modalités d'associations différentes au sein du même espace sémantique, suggérant que les locuteurs attribuent des charges sémantiques particulières à ce même mot. Les profils d'Arlette Laguiller et de Nicolas Sarkozy se distinguent face aux profils aplatis de Buffet, Royal, Bayrou et Le Pen. Or Mayaffre (2008) avait pressenti avec des outils textométriques classiques (calcul des spécificités) qu'il s'agissait bien d'une thématique commune aux deux candidats (*versus* les autres) par delà leur engagement politique différent.

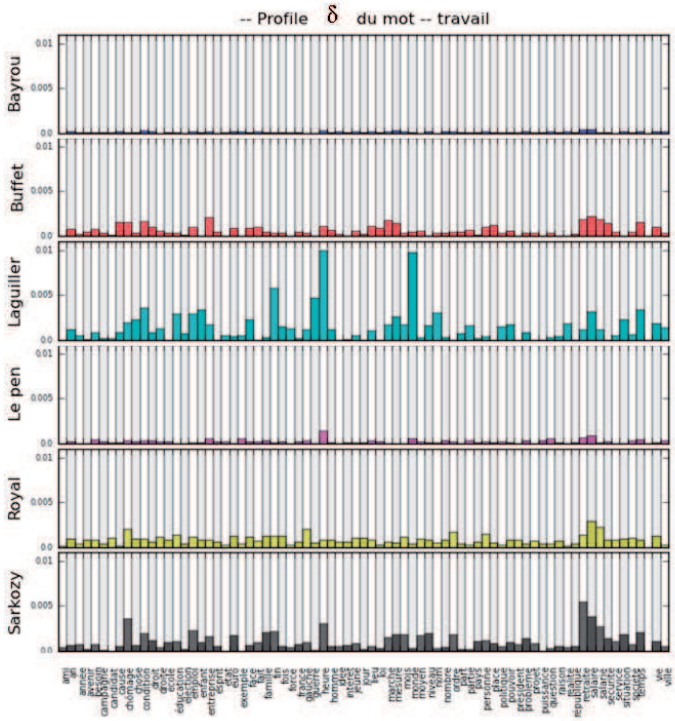

Figure 2 - Disponibilité δ du mot 'travail' chez les 6 candidats présidentiels pour les 75 mots de leur vocabulaire commun, suggérant des choix discursifs différents chargeant le mot 'travail' sémantiquement de façon différente selon le candidat. L'ordre des mots est le même (alphabétique) pour tous les profils, ainsi que la calibration de l'axe des ordonnées.

De plus, si l'on compare désormais entre eux Laguiller et Sarkozy, on constate que les valeurs les plus fortes ne concernent pas les mêmes mots : 'retraite' et 'salaire' pour Sarkozy, 'monde' et 'heure' pour Laguiller. Le retour aux textes permet alors de confirmer qu'autour d'une préoccupation commune organisée autour du mot 'travail', les deux candidats ne thématisent pas le sujet de la même manière. Pour Sarkozy, c'est historiquement le 'travailler plus pour gagner plus' qui domine avec la nécessité de retarder l'âge de départ à la 'retraite' et l'espoir

de disposer d'un meilleur 'salaire'. Pour Laguiller, c'est l'organisation même du travail (le 'monde' du travail) et du problème des conditions sociales autour de la durée légale du travail ('heure') qui occupent le discours. Ainsi, à titre illustratif, nous trouvons respectivement des extraits caricaturaux comme suit :

> « Une politique qui cherche à créer du TRAVAIL au lieu de chercher à le partager. Je veux que ceux qui veulent travailler plus pour gagner plus puissent le faire […]. Je veux que les retraités soient libres de travailler et de cumuler leur RETRAITE avec un SALAIRE.» (Sarkozy, 28 mars 2007, meeting de Lille)

> « Exonérer les patrons de charges sociales et fiscales sur les HEURES supplémentaires, comme le promet Sarkozy, est non seulement leur faire un beau cadeau, mais c'est une façon de les pousser à faire crever au TRAVAIL leurs travailleurs et de ne pas embaucher. Si la série de suicides à Renault Guyancourt et puis à Peugeot Charleville ont tant ému le MONDE du TRAVAIL, c'est parce que chacun ressent que les conditions de TRAVAIL sont pour quelque chose dans ces actes désespérés. » (Laguiller, 23 février 2007, meeting du Mans)

L'asymétrisation des profils co-occurrentiels en termes de disponibilité aurait donc une valeur de prédictibilité thématique.

3.3. Coordination des mots dans un seul espace normé

Chaque mot décline son énergie et sa disponibilité en profils pour chacun des candidats. Afin de constituer une vue synthétique des comportements de *tous les mots à la fois au regard de leur énergie et de leur disponibilité*, nous avons implémenté la proposition de Bonneau [2012 – à paraître] de résumer les profils ε / δ de chaque mot en deux coordonnées $\left\|\overline{\varepsilon}\right\|$ / $\left\|\overline{\delta}\right\|$ calculées comme la norme de chacun des profils ε et δ, respectivement :

$$\begin{cases} \left\|\overline{\varepsilon}(A)\right\| = \sum_{\forall B} \left(\varepsilon(A)_{in\ AB}\right)^2 \\ \left\|\overline{\delta}(A)\right\| = \sum_{\forall B} \left(\delta(A)_{in\ AB}\right)^2 \end{cases} \quad (6)$$

La norme mesure la dispersion des énergies (respectivement, des disponibilités) dans l'espace couvert par le profil. Etant donné que tous les profils sont de même longueur, $\left\|\overline{\varepsilon}\right\|$ et $\left\|\overline{\delta}\right\|$ sont directement comparables. Chaque mot est alors caractérisé par 2 coordonnées que l'on peut projeter dans un espace normé (Fig. 3).

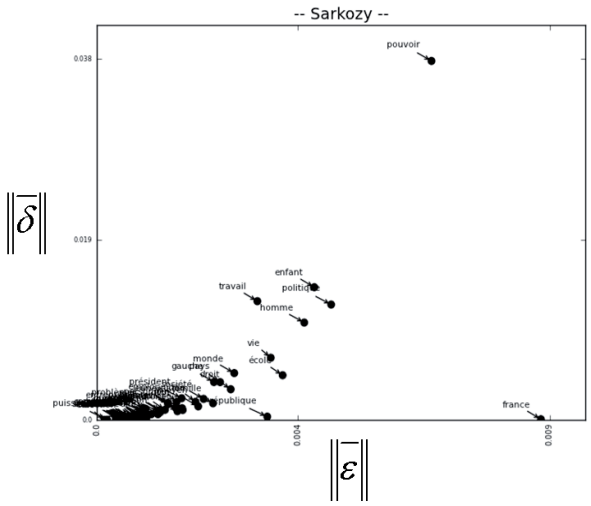

Figure 3 - Projection du vocabulaire commun aux 6 candidats dans l'espace normé $\|\overline{\delta}\| = f\left(\|\overline{\varepsilon}\|\right)$ *pour le candidat Sarkozy. Chaque point correspond à un mot.*

La majorité des mots affiche une norme énergie et une norme disponibilité faibles, et est agglomérée prés de l'origine. Ces mots se trouvent être des mots sans charge sémantico-politique évidente comme 'fait', 'chose', 'cause', 'campagne', 'besoin'. En analyse du discours, on aurait tendance à les exclure de l'analyse en tant que mots outils du discours politique. Certains mots en revanche se détachent soit du point de vue de leur énergie soit du point de vue de leur disponibilité et se retrouvent plus excentrés (vers le haut ou la droite) sur le graphique. Selon l'interprétation proposée en 1.3, une norme disponibilité $\|\overline{\delta}\|$ élevée caractériserait des mots fortement chargés sémantiquement dans le cadre discursif considéré, ce qui est cohérent avec des mots comme 'pouvoir', ou encore 'travail' (surtout chez le candidat Sarkozy ou chez Laguiller (Fig. 4), 'école' ou 'enfant'. Toutefois, les mêmes mots peuvent avoir des comportements différents selon le candidat envisagé. Chez la candidate Laguiller, par exemple, ce sont des mots comme 'entreprise', 'emploi' et 'salarié' qui se retrouvent dans cette région de forte $\|\overline{\delta}\|$, et chez la candidate Royal, ce sont les mots 'emploi', 'jeune', 'travail' et 'pays' qui se différencient de la sorte (Fig. 4).

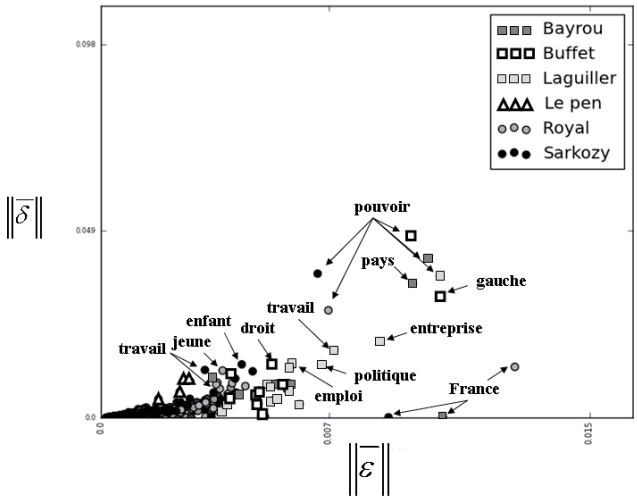

Figure 4 - Projection du vocabulaire commun aux 6 candidats dans l'espace normé $\|\overline{\delta}\| = f\left(\|\overline{\varepsilon}\|\right)$ *pour tous les candidats. Chaque point correspond à un mot.*

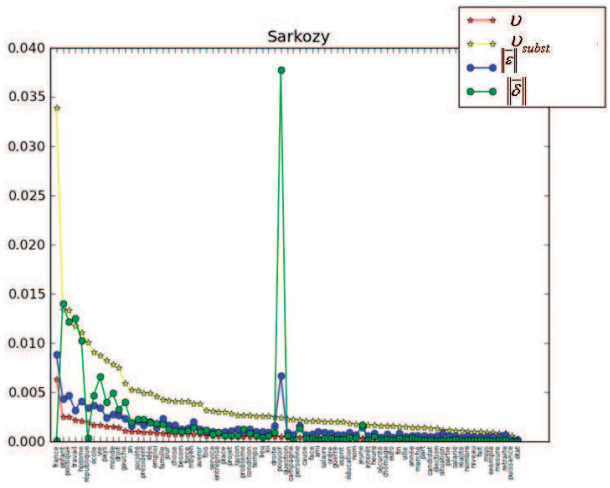

Figure 5 - Variation du couple $\|\overline{\varepsilon}\|$ / $\|\overline{\delta}\|$ *en fonction de la fréquence relative d'usage du mot dans le corpus du candidat Sarkozy. Les mots sont triés par fréquence décroissante.*

Le comportement en énergie du mot 'France', qui se trouve être un mot très fréquent, tous textes confondus, a attiré notre attention sur un possible effet de fréquence dans la répartition des mots dans l'espace normé $\|\overline{\delta}\| = f\left(\|\overline{\varepsilon}\|\right)$. La Fig. 5 montre que les variations de $\|\overline{\delta}\|$ et surtout celles

de $\left\|\overline{\varepsilon}\right\|$ sont, en tendance, sensibles à la fréquence relative du mot considéré. Toutefois, l'effet de fréquence, par ailleurs attendu puisque la première stratégie discursive de thématisation concerne la fréquence d'utilisation de tel ou tel mot, n'est pas pour autant suffisant à expliquer la répartition des mots dans l'espace considéré. Ce que l'on observe sur de tels graphiques semble bien relever de la coloration sémantique différente d'un mot à l'autre et d'un locuteur à l'autre. Le tout rejoint notre conclusion, en 3.2 : *une approche asymétrique de la co-occurrence permet de diagnostiquer des éléments de thématisation*.

3.4. Le rendement discursif : un seul profil pour saisir l'asymétrie co-occurrentielle

La représentation du comportement sémantique des mots grâce à leurs coordonnées $\left\|\overline{\varepsilon}\right\|$ / $\left\|\overline{\delta}\right\|$ est certes synthétique, mais elle se fait au détriment de l'information spécifique des modes d'associations mutuelles que chaque mot entretient avec chaque autre. Afin de restituer cette dimension spécifiante du profil tout en conservant une articulation synthétique des apports respectifs de chaque pôle à la structuration sémantique de la co-occurrence, nous introduisons la notion de *rendement discursif* de A (dans sa relation à tout B), noté $r_d(A)$:

$$\forall B, r_{\mathrm{d}}(A) = \delta(A)_{in\ AB} - \varepsilon(A)_{in\ AB} \quad (7)$$

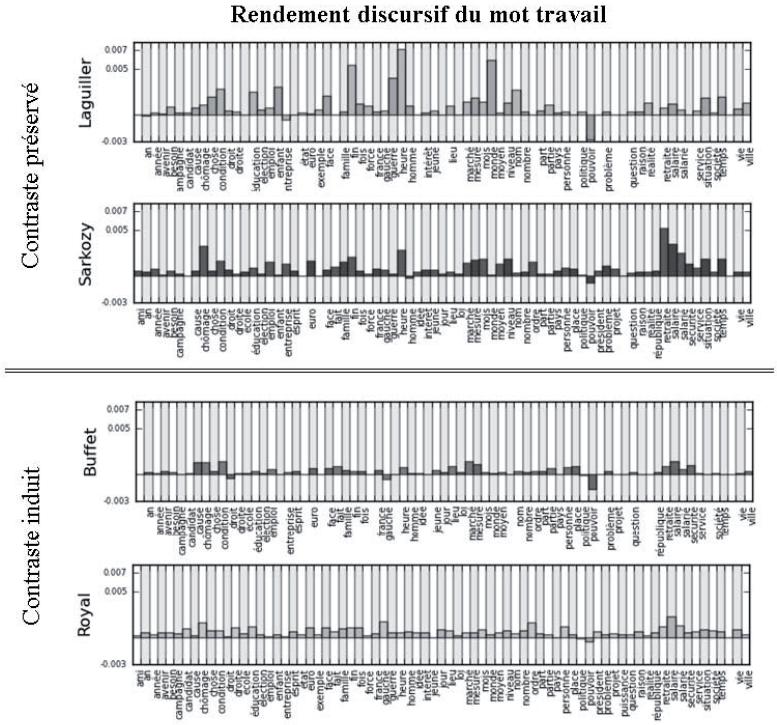

Figure 6 - Rendement discursif du mot 'travail' pour les candidats Laguiller, Sarkozy, Buffet et Royal.

La Fig.6 montre les profils de rendement discursif du mot 'travail' chez deux groupes de candidats. Le premier groupe est composé de Laguiller et Sarkozy, qui manifestaient dans la

Fig.2 des usages différenciés du mot ‘travail’ l’un par rapport à l’autre. Ce contraste d’usage est conservé dans les profils de rendement discursif, qui sont peu altérés par rapport aux seuls profils de disponibilité. En ce qui concerne le deuxième groupe, composé de Buffet et de Royal, *les profils de rendement discursif introduisent un contraste d’usage qui n’était pas très apparent à partir des seuls profils de disponibilité* du mot ‘travail’ chez ces candidates.

4. Asymétrie co-occurrentielle et thématisation

L’ensemble des observations reportées en section 3 appuie une interprétation de la disponibilité d’un mot en rapport avec son comportement en discours, alors que l’énergie saisit quelque chose d’un autre ordre - putativement son comportement en langue (Luong *et al.*, 2010) même si cela n’est pas soutenu directement par nos analyses. Si la disponibilité mesure ou identifie effectivement la charge sémantique investie dans un mot par choix discursif de celui qui le dit, elle est alors en mesure de diagnostiquer les thèmes développés dans un discours. Nous nous proposons, dans cette dernière section, de tester la capacité relative de la disponibilité et du rendement discursif à identifier et circonscrire les usages thématiques d’un mot, comme première étape à une utilisation plus généralisée d’extraction probabiliste de thèmes discursifs.

4.1. Méthode : extension du champ d’association sémantique

Dans la section 3, nos analyses ont porté uniquement sur les substantifs *partagés* par les 6 candidats parmi les substantifs *les plus fréquents du corpus total*. Cette double contrainte s’est avérée très sélective, construisant une liste contenant un grand nombre de substantifs outils du vocabulaire politique (‘question’, ‘problème’, ‘fait’, ‘cause’, etc.) et peu de substantifs pouvant donner matière à un choix discursif, soit du fait que ceux-ci font l’objet d’un usage moins fréquent au niveau du corpus total, soit que leur usage n’est pas partagé par tous les candidats.

Afin d’étudier plus finement la capacité des mesures de disponibilité et de rendement discursif à identifier les thématiques déployées dans un discours, nous avons procédé à une révision des contraintes de sélection des associés étudiés, de sorte à étendre le champ de l’interprétation thématique. Ayant identifié, *pour chaque candidat*, les 150 substantifs les plus fréquemment utilisés, nous constituons une liste étendue à partir de *l’union* de ces 6 sous-listes, qui comprend 374 mots. Ce choix est arbitraire, et cette liste aurait pu être constituée différemment (sur la base des spécificités statistiques, par exemple), mais notre propos ici n’est pas encore de constituer une approche généralisée de la thématisation, mais est plutôt prospectif quant à l’utilité de certaines mesures de l’asymétrie co-occurrentielle à la caractériser.

4.2. Disponibilité et diagnostic d’une thématisation différentielle du mot ‘immigration’

L’échantillonnage utilisé assure une représentation égale aux 6 candidats. Toutefois, un mot comme ‘immigration’, qui est présent chez tous, se trouvera associé de façon différente, tant en termes de champ d’association que de force d’association (Tab. 1) par les différents candidats.

Dans le corpus l’immigration apparaît comme une thématique de droite, avec un champ d’associés plus restreint chez les candidats de gauche, surtout à l’extrême du spectre. Pour les mots non outils du discours politique, Buffet associe, en disponibilité, le mot ‘immigration’ avec ‘justice’ et ‘force’ quand Laguiller l’associe à ‘classe’ et à … ‘Sarkozy’.

Candidat	Nombre d'associés tel que δ (immigration) ≠ 0	$\overline{\delta}$	$\sigma(\delta)$	cv(δ)
Laguiller	6	4.121e-07	3.689e-07	89.5%
Buffet	6	2.362e-06	1.749e-06	74%
Royal	46	5.201e-06	6.391e-06	122.8%
Bayrou	169	4.412e-05	8.072e-05	182.9%
Sarkozy	127	5.501e-05	8.963e-05	162.9%
Le Pen	142	2.461e-04	2.089e-04	84.8%

Table 1 – Nombre d'associés pour lesquels le mot 'immigration' présente une disponibilité δ(immigration) non nulle et distribution des valeurs de δ(immigration) sur ce profil cooccurentiel restreint (moyenne $\overline{\delta}$, écart-type $\sigma(\delta)$ et coefficient de variation $c_v(\delta)$)

A droite, Bayrou, qui se distingue par le plus grand nombre d'associations en disponibilité le fait aussi avec le plus fort coefficient de variation (182.9%). Avec un profil de variation d'apparence semblable à celui de Sarkozy (en moyenne, en variance, et en forme du profil), les co-occurrents impliqués thématisent de façon différente la question de l'immigration chez les 2 candidats (Fig. 7) : même si Sarkozy le fait avec de faibles disponibilités, il associe à 'immigration' les mots 'peur', 'voyou', 'communautarisme', ou encore 'civilisation', 'échec' et 'chômage', termes que Bayrou n'utilise pas pour construire la charge sémantique du ce même mot.

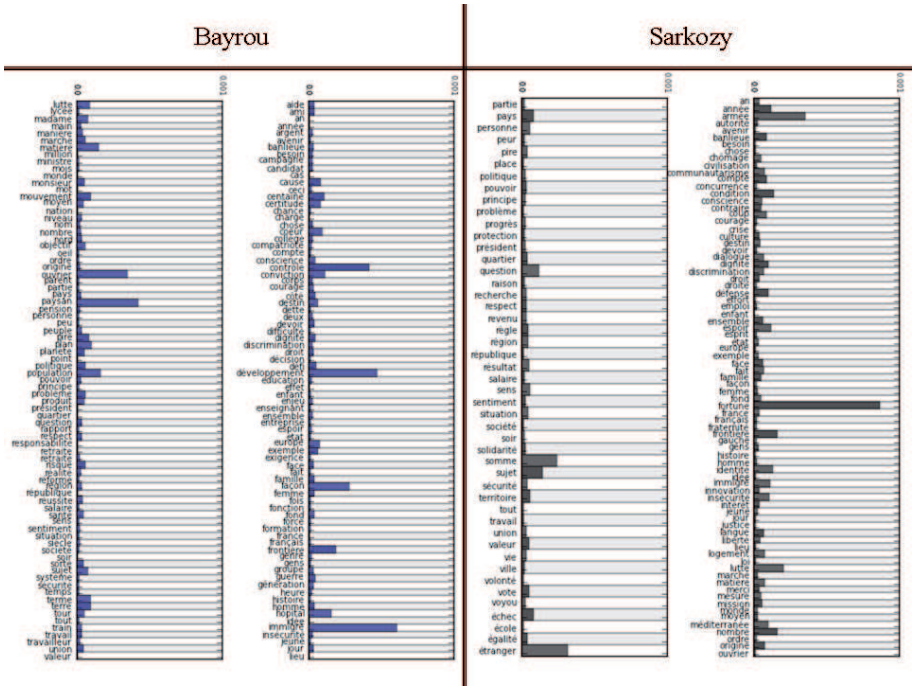

Figure 7 – Profils thématiques étendus, en disponibilité, du mot 'immigration, chez Bayrou et Sarkozy.

En ce qui concerne le discours de Le Pen, le mot ‘immigration’ semble surdéterminé sémantiquement : non seulement il recrute un grand nombre d’associés pour construire sa charge sémantique, mais en plus, chacun de ceux-ci contribue avec une énergie (c’est-à-dire une disponibilité du mot ‘immigration’) plus forte en moyenne que chez les autres candidats (Tab. 1).

4.3. Répartition de la charge sémantique et modalités de thématisation

La réinterprétation des profils co-occurrentiels en termes de rendement discursif accentue les observations faites pour la disponibilité. En effet, les candidats de gauche ont (à l’exception du mot ‘méditerranée’ chez Royal) un rendement négatif : pour chaque co-occurrence profilée, l’énergie de ‘immigration’ est supérieure à sa disponibilité. Dit autrement, c’est la disponibilité du co-occurrent de ‘immigration’ qui se trouve, quasi systématiquement être supérieure à celle de ‘immigration’. A droite, Bayrou et Sarkozy ont un comportement plus nuancé, qui confirme les tendances de thématisation observées sur les profils de disponibilité. Seul Le Pen se trouve être quasi systématiquement dans des valeurs positives de rendement discursif pour le mot ‘immigration’ (Fig. 8).

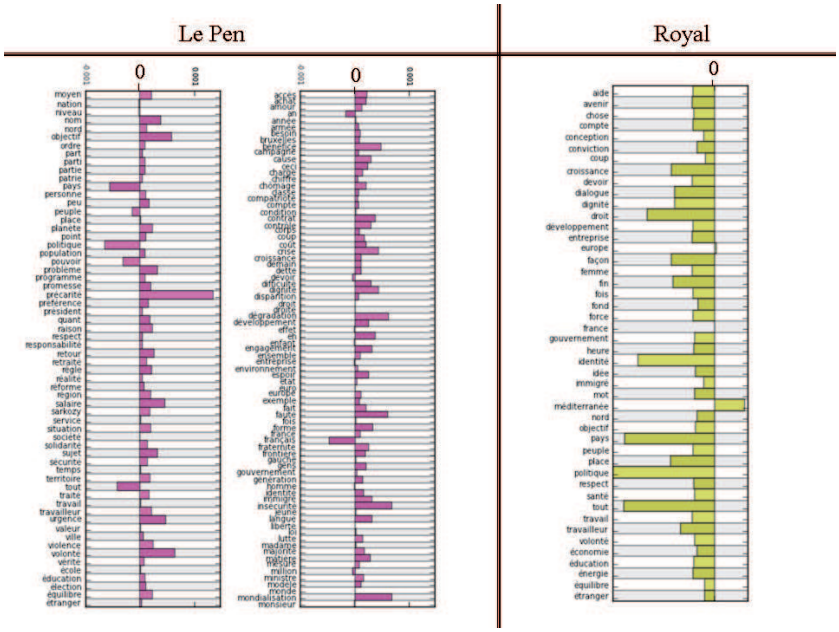

Figure 8 – Profils thématiques étendus, en rendement discursif, du mot ‘immigration, chez Royal et Le Pen. Royal illustre la tendance observée chez tous les candidats de gauche, alors que le Pen se démarque à droite par des rendements discursifs essentiellement positifs.

Si la disponibilité apporte une information essentielle sur la thématisation de tel ou tel mot, il nous semble, à partir de cet exemple, de celui présenté en 3.4, et sur d’autres exemples que nous ne pouvons exposer ici, que le rendement discursif, en articulant la disponibilité de l’un et de l’autre pôle de la co-occurrence, apporte une information supplémentaire sur les modes de thématisation adoptés par le locuteur.

5. Conclusion

La disponibilité nous semble en mesure de saisir la notion de charge sémantique, et est, à ce titre, utile pour le diagnostic de la thématisation d'un mot. Toutefois, il nous semble essentiel de capturer l'asymétrie co-occurrentielle dans son ensemble en combinant la disponibilité et l'énergie ; cette dernière étant en définitive une disponibilité, mais vue de l'autre pôle de la co-occurrence. La notion de rendement discursif que nous proposons ici semble pouvoir nuancer des modalités de thématisation d'un mot donné que la disponibilité seule ne parvient pas à diagnostiquer.

Cette étude reste très largement exploratoire, et mériterait d'être systématisée afin de déterminer le potentiel réel du triplet énergie/disponibilité/rendement discursif pour établir les thèmes du discours ou, pour un même mot, ses modes de thématisation dans un corpus contrastif. On identifierait alors les spécificités discursives entre textes ou locuteurs en replaçant le mot à la fois dans son profil co-occurrentiel et dans l'asymétrie qui le lie à chacun de ses associés.

Bibliographie

J. Bonneau (à paraître en 2012), *Approche mathématique de la textualité. Application au corpus Mendès France (1924-1960)*, thèse de doctorat, Université de Nice.

K. Church and P. Hanks (1991). "Word Association Norms, Mutual Information and Lexicography", *Computational Linguistics*, Vol 16 : 1, pp. 22-29.

J.R. Firth (1957). *Papers in linguistics, 1934–1951*. Oxford : Oxford University Press.

S. Heiden (2004). "Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex", *JADT 2004 - Le poids des mots – Actes des 7es Journées Internationales d'Analyse Statistique des Données Textuelles*, édité par G. Purnelle, C. Fairon, A. Dister. Presses Universitaires de Louvain, pp. 577-588.

X. Luong, E Brunet, D. Longrée, D. Mayaffre, S. Mellet et C. Poudat (2010). "La cooccurrence, une relation asymétrique ?". JADT 2010 - Statistical Analysis of Textual Data Proceedings of the 10th International Conference, 9-11 June 2010 - Sapienza University of Rome, édité par S. Bolasco, I. Chiari, L. Giuliano. Milan : Edizioni Universitarie di Lettere Economia Diritto, pp. 321-331.

W. Martinez (2003). Contribution à une méthodologie de l'anlayse des cooccurrences lexicales multiples dans les corpus textuels. Thèse de doctorat en Sciences du Langage, Université de la Sorbonne nouvelle – Paris 3, sous la direction d'André Salem, Paris.

D. Mayaffre (2008). "Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence", in Serge Heiden et Bénédicte Pincemin (éds.), *JADT 2008, 9es journées internationales d'analyse statistique des données textuelles*, Lyon, Pul, vol. 2, pp. 811-822.

F. Rastier (2011). La mesure et le grain. Sémantique de corpus. Paris : Champion.

J.-M. Viprey (1997). Dynamique du vocabulaire des Fleurs du mal. Paris : Champion